Biotechnology Progress

Group screening for rare events based on unreduced incomplete block designs

Journal:	*Biotechnology Progress*
Manuscript ID	BTPR-18-0236.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	06-Dec-2018
Complete List of Authors:	Eskridge, Kent Gilmour, Steve; King's College London School of Natural and Mathematical Sciences, Mathematics Posadas, Luis; University of Nebraska, Lincoln, Agronomy and Horticulture
Keywords:	Group testing, screening, pooling, prevalence, : Group testing, multistage, pooling, prevalence

Group screening for rare events based on incomplete block designs*

K. M. Eskridge[1], S. G. Gilmour[2] and L. G. Posadas[3]

1. *Department of Statistics, Univ. of Nebraska, Lincoln, NE, USA*
2. *Department of Mathematics, King's College, University of London, UK*
3. *Department of Agronomy and Horticulture, Univ. of Nebraska, Lincoln, NE, USA*

Correspondence about this article should be addressed to K. M. Eskridge @ keskridge1@unl.edu

Abstract

Fields such as, diagnostic testing, biotherapeutics, drug development and toxicology among others, center on the premise of searching through many specimens for a rare event. Scientists in the business of "searching for a needle in a haystack" may greatly benefit from the use of group screening design strategies. Group screening, where specimens are composited into pools with each pool being tested for the presence of the event, can be much more cost efficient than testing each individual specimen. A number of group screening designs have been proposed in the literature. Incomplete block screening designs are described here and compared with other group screening designs. It is shown under certain conditions, that incomplete block screening designs can provide nearly a 90% cost saving compared to other group screening designs such as when prevalence is 0.001 and screening 3876 specimens with an ICB-sequential design vs a Dorfman design. In other cases, previous group screening designs are shown to be most efficient. Overall, when prevalence is small (≤0.05) group screening designs are shown to be quite cost effective at screening a large number of specimens and in general there is no one design that is best in all situations.

Keywords: Group testing, multistage, pooling, prevalence

* This manuscript is a contribution of the Univ. of Nebraska, Agric. Res. Div., supported in part by funds provided through the Hatch Act, USDA.
Email address: keskridge1@unl.edu

Introduction

The development of new technologies that allow for rapid and multiplexed diagnosis and screening tests represent promising methods for quickly detecting animal and plant infectious diseases, tumor-derived biomarkers, plant seed genetic purity, food safety, environmental pest management and other types of molecular polymorphisms with important implications ranging from agriculture to biosecurity. (*1*) - (*2*) However, the incorporation of powerful experimental design strategies to further increase the potential of many high-throughput molecular platforms now available has received little attention. Such is the case of group screening designs, which can have a significant impact in the cost and efficiency of experiments where a positive event is rare and samples sizes are large.

To illustrate the usefulness of a group screening design, assume a plant geneticist wants to evaluate 455 plant genotypes for the presence of a particular transposition event where fewer than 1% are expected to contain the event. Expensive and time consuming DNA extractions and PCR-based identification techniques must be used to identify the plant carrying the transposition event. In order to reduce costs, the geneticist would like to pool the DNA of groups of plants where extractions are conducted only on the pools. Plants in any pool that does not contain the particular transposition event will be eliminated from further consideration. Plants in positive pools could either be tested individually or in smaller pools. Using such a group screening strategy will generally result in fewer tests needed to identify all positive plants compared to separately testing all 455 plants. (*3*)

Two different group screening designs illustrate the strategy for this particular study. A simple approach, assuming sufficient test sensitivity, would be to composite all 455 plants in one pool and if the pool tested negative, clear all 455 plants of containing the event, while

if the pool tested positive, test all the 455 individually. However, the expected number of tests would likely be too large except for cases where the probability of an event was much smaller than 1%. Another approach, proposed by Dorfman (*4*) would be to divide the 455 specimens into *m* pools, (say for example 35 pools of 13 plants each) where each specimen in a pool would be retested if the pool tested positive. Dorfman (*4*) demonstrated that this scheme could result in cost savings up to 80% over testing all specimens separately

In general, group screening of *b* specimens (plants in the example) proceeds by assigning specimens to pools using some design, clearing all specimens in negative pools and further testing specimens in positive pools either separately or again in batch . Phatarfod and Sudbury (*5*) proposed arranging specimens in a square array with $b=n^2$ and pooling each of *n* rows and *n* columns, giving *2n* pools, which improved efficiency over the Dorfman approach in some situations while Hudgen and Kim (*6*) considered optimal configurations of square arrays. Sudbury (*7*) compared efficiencies between two-dimensional arrays and selection designs and found selection designs to be at least as efficient as two-way arrays for a range of different batch sizes. Berger et al. (*8*) proposed higher-way array pooling and demonstrated that these methods could further improve efficiency. Kim and Hudgens (*9*) showed that three-dimensional array based methods in the presence of test error are more precise than two-dimensional arrays and that four-way and higher arrays do not lead to large gains in efficiency. A number of multi-stage and sequential schemes have been proposed where most have used the Dorfman group screening design as an initial design (*10*)- (*11*).

In this work, we propose using unreduced incomplete block designs for developing group screening designs and show that under some conditions, these designs perform better

than previously proposed methods when the probability that an individual specimen is positive is small (≤ 0.05). Sudbury (*7*) considered unreduced incomplete block group screening designs, which he termed selection designs, but did not consider expected number of tests to determine all positive specimens, compare them with designs other than 2-way arrays and did not consider sequential applications of these designs. Other types of incomplete block designs have been proposed for group screening designs but generally for different purposes. Redman and King (*12*) proposed using balanced and partially balanced incomplete block designs for group screening when quantitative response was available for each pool. Bush et al. (*13*) proposed using special types of incomplete block designs called *t*-complete designs useful when the number of defectives is known before experimentation which is unlikely in many applications. Balding and Torney (*14*) proposed using various types of incomplete block designs in a non-adaptive setting where specimens were not retested. Du and Hwang (*15*) described the theory and general properties of using incomplete block designs as group screening designs. Here, we first show how to construct an unreduced incomplete block group screening (ICB) design and in general how these designs are based on unreduced incomplete block designs, then describe the data analysis and how expected sample sizes may be determined. We then compare several different group screening designs with the designs proposed in this paper and finish with some discussion.

Design Construction

The fundamental idea of an ICB group screening design is to (1) determine all possible ways an individual specimen can be assigned to k out of t pools and (2) add a single

pool that contains all specimens. This design results in each specimen being tested in $k+1$ pools. For example, assume $10 = \binom{5}{2}$ specimens are to be tested with a total of 6 (5+1) pools where each specimen is to be present in 3 (2+1) pools. First, determine all possible ways that a specimen can be assigned to 2 of 5 pools resulting in Pools 1 through 5 of Table 1. Then add a final pool with all specimens which is the 6th pool in Table 1. Finally, randomize the specimen numbers in Table 1 to actual specimens. In Table 1, specimen 1 is in pools 1, 2 and 6 while specimen 9 is in pools 3, 5 and 6. In the data analysis, the sequence of positive pools can identify which specimen is positive. For example, if pools 1, 2 and 6 are positive, then specimen one is uniquely identified as positive while all others are negative. With this particular design, up to two positive specimens can be detected although not uniquely. For example, if all pools are positive except 3, then either (i) specimens 1 and 10 or (ii) specimens 4 and 6 or (iii) specimens 3 and 7 must be positive, while the others are negative.

In general, assume n pools are to be used to screen $b = \binom{t}{k}$ specimens (b n) for the presence of a particular attribute where $t=n-1$ and each specimen is assigned to $k+1$ pools. Each pool, excluding the final pool containing all specimens, will contain $\binom{t-1}{k-1}$ specimens. Also, assume that within the population, the chances that any individual specimen is positive is quite small (≤ 0.05) and that there are no dilution, interaction effects or errors in the testing technique. Then a negative pool response eliminates those specimens in the pool from further testing. This type of group screening design can be shown to be obtained from an unreduced incomplete block design (ICB) where the b

specimens form the incomplete blocks and the first *t-1* pools represent the treatments where there are *k* treatments in each block. For example the first 5 pools in Table 1 are based on an unreduced incomplete block design with 5= *t-1* treatments in *10*= *b* incomplete blocks of size 2=*k,* where the 6th pool is augmented to the other 5 pools. Each of the first 5 pools will contain $\binom{5-1}{2-1} = 4$ specimens.

This construction method is general in that the approach may be used with any unreduced incomplete block design (ICB) (ie. any *t* and *k* where $k<t$) however, the approach is limited to the number of test specimens where $b=\binom{t}{k}$. Any incomplete block design with *t* treatments and *k* units per block (bib or not, unreduced or not) could be used to identify a group screening design, however the properties of these designs have not be considered. As an example, modifying the design in Table 1 to screen 8 instead of 10 specimens by eliminating specimens (blocks) 9 and 10 results in an ICB screening design based on an imbalanced incomplete block design which has unknown properties regarding expected sample size (see below).

Data Analysis

The resulting data from an ICB group screening experiment will be a plus (+) or minus (–) for every pool where a + for a pool indicates that at least one of the specimens in that pool is positive and a – if the pool contained no positive specimens. Specifically the results will be a *n* vector of pluses (+) and minuses (–) and data analysis is based on this response vector. The idea is to retest those specimens that could be contributing to the particular response vector or alternatively, eliminate those specimens from further testing which could not.

Assume n pools are to be used to screen $b=\binom{t}{k}$ specimens. The number of positive specimens will depend on the particular response vector.

All pools are negative. If all elements are negative in the response vector, then no specimens are positive and no more testing is required. Note that it is not possible to have between one and k positive elements in the response vector, since if a specimen is positive, at least $k+1$ positive values will result since each specimen is in $k+1$ pools. If there are between one and k positives, then the assumptions of no error, dilution or interaction effects may be incorrect.

Exactly one negative. If there is exactly one – sign present in the response vector, then there are $\binom{t-1}{k}$ different specimens (or blocks) that could have contributed to the t positive signs in the response vector and these specimens should be retested. Specimens are retested if they are in any of the t pools that tested positive, or alternatively, any specimen is eliminated from further testing if it is only contained in those pools that test negative.

Exactly two negatives. If there are exactly two – signs in the response vector, then $\binom{t-2}{k}$ specimens (or blocks) may be contributing to the response vector. Specimens are retested if they are in any one of the $t-1$ pools that tested positive.

General Case. If a group screening design is based on an unreduced incomplete block design (t,k) and if there are p – signs ($0 < p \leq t - k$) in the response vector, then $\binom{t-p}{k}$ specimens could be contributing to the response and should be retested. Specimens

are retested if they are in any of the *t-p+1* pools that tested positive, or alternatively, any specimen is eliminated from further testing if it is contained in all pools that test negative.

Expected Sample Size

An ICB group screening design will be useful if it reduces the number of pools (or samples) needed compared to other group screening designs. To determine the expected sample size (ie, the expected number of retests plus the initial number of pools), assume that (i) specimens are independent and randomly sampled from the population of interest and the probability that any particular specimen is negative is q, (ii) there are $b = \binom{t}{k}$ specimens and n pools (or $t = n-1$), where each specimen will be tested in $k+1$ pools and (iii) one of two different follow-up schemes will be used with the specimens to be retested: (a) either tested separately or (b) all retested using additional ICB group screening designs.

Retested specimens tested separately. To determine the expected number of retests, we need the probability distribution of the number of retests:

Number of retests	0	$\binom{k}{k}$	$\binom{t-(t-k-1)}{k}$	$\binom{t-(t-k-2)}{k}$	…	$\binom{t-1}{k}$	$\binom{t}{k}$
Probability	q^b	$\binom{b}{1}q^{b-1}(1-q)$	$f_1(q)$	$f_2(q)$	…	$f_{t-k-1}(q)$	$f_{t-k}(q)$

The expected number of retests is computed in the usual way: *E(no. of retests)* = ∑*(no. of retests)(probability).*

The idea of this probability distribution is that the binomial distribution of the number of positive specimens out of *b* specimens will be 'broken-up' into the needed probabilities above. Here, $f_u(q) = C_{u1}\, q^{b-2}(1-q)^2 + C_{u2}\, q^{b-3}(1-q)^3 + C_{u3}\, q^{b-4}(1-q)^4 \ldots$ *(u=1,...,t-k),*

$$C_{u1} = \binom{k}{k-u}\binom{t-k}{u}\binom{t}{k} \div 2 \quad (u=1,\ldots,k)$$

$$C_{uv} = \sum_{j=1}^{t-k} C_{j,v-1}\left[\binom{k+j}{j+k-u}\binom{t-(k+j)}{k-(j+k-u)} - vI(j=u)\right] \div (v+1) \qquad (v=2,\ldots.b-1)$$

where C_{u1} and/or C_{uv} are defined as 0 if any combination $\binom{x}{z}$ in their definition has either $x<0$, $z<0$, or $x<z$. Also, by definition of the binomial distribution $\sum_{u=1}^{t-k} C_{uv} = \binom{b}{v+1}$ for $v=1,\ldots,b-1$. See Appendix for proof.

As an example, the design with b=10 (t=5,k=2) has the following probability distribution of the number of retests and expected number of retests.

Number of retests	Probability
$\binom{2}{2} = 1$ 0	q^{10}
	$\binom{10}{1} q^9(1-q)$
$\binom{3}{2} = 3$	$30q^8(1-q)^2 + 10q^7(1-q)^3$

$\binom{4}{2} = 6$ $\quad 15\,q^8(1-q)^2 + 80q^7(1-q)^3 + 75\,q^6(1-q)^4 + 30\,q^5(1-q)^5 + 5q^4(1-q)^6$

$\binom{5}{2} = 10$ $\quad 30q^7(1-q)^3 + 135\,q^6(1-q)^4 + 222\,q^5(1-q)^5 + 205q^4(1-q)^6$

$+ 120q^3(1-q)^7 + 45q^2(1-q)^8 + 10\,q^1(1-q)^9 + (1-q)^{10}$

Using the definition of expected value and the probability distribution in the display above, the expected number of retests when q=0.999 can be shown to be 0.01 and the expected sample size would be 6.01.

Retested specimens tested in follow-up ICB group screening designs. Using concepts from unreduced incomplete block designs, the number of specimens to be retested given any particular response vector from the analysis of the initial design, may be expressed as a combination. If the initial design has parameters (t,k) and there are p – signs (p 0) in the response vector, $\binom{t-p}{k}$ specimens must be retested, where the specific specimens to be retested are those that are in the pools that tested positive. These $\binom{t-p}{k}$ specimens can be tested in a follow-up ICB group screening design with 'new' parameters $(t\text{-}p, k)$ (or $(t\text{-}p, t\text{-}p\text{-}k)$ by symmetry of binomial coefficients). This follow-up design is likely to have a much smaller expected sample size than testing all of the $\binom{t-p}{k}$ specimens.

For example, if the initial design has 462 specimens ($t=11, k=5$) and there are $p=2$ – signs in the response vector, there will be 126 specimens to be retested. These 126 specimens are then tested using a $t=9,k=5$ ICB group screening design. Depending on the size of q, this approach may result in a considerable savings over separately testing all 126 specimens. If q=0.999, then the expected number of retests is 0.37 and the expected total number of tests

for the *t=9, k=5* design is 10.37 vs 126. This testing with follow-up ICB group screening designs may be repeated successively in a multi-stage approach which could reduce the expected sample size even further. For example, if the initial design is (*t=11,k=5)* and *p=2* – signs result, a second stage follow-up design with *t=9* and *k=5* could be used to test the remaining 126 specimens. If with the second stage, 2 – signs result, then the 21 specimens to be retested would be evaluated with a *t=7, k=5* design. If the third stage resulted in a single – sign, the just one specimen would be retested for a total of 11+9+7+1 = 28 tests in all.

To use the proposed ICB designs, a researcher would first establish the number of specimens to be screened and then construct the appropriate number of pools and analyze results as described above. For example, assume a researcher has 1365 specimens to evaluate, that q=0.999 in the population and wants to minimize the expected number of required pools. Given Table 2, use of an ICB design shows considerable cost savings in terms of the expected number of pools necessary (eg., 18 fixed ICB (ICB-F); vs 86 Dorfman; 75 Two-way; 40 Three-way; and 1017 single pool).

For a more realistic example with parameters not given in Table 2, consider a toxicology project in protein therapeutics, where a cell-line derived through cloning is used for the purposes of synthesizing therapeutic protein material for toxicology testing and, subsequently, for clinical studies. Assume the identification of the "right" clone is key to this research and that there is a large population of stably transfected cell lines, each expressing a different antibody type. The goal is to identify the antibody that best targets a very specific epitope (i.e., selective targeting). In this case, this is a line that can be identified through a binary response, e.g., above *vs* below a response threshold. Assume that

this is a rare event, e.g., it occurs 1 out of every 1,000 cell lines tested (q=0.999), and that there are no dilution, interaction effects or errors in the testing technique. Suppose the researcher has 4560 cell-lines available for testing. Instead of screening each clone individually and investing resources into at least 48 96-well ELISA plates and reagents, the researcher could use the scenario described in Figure 1.

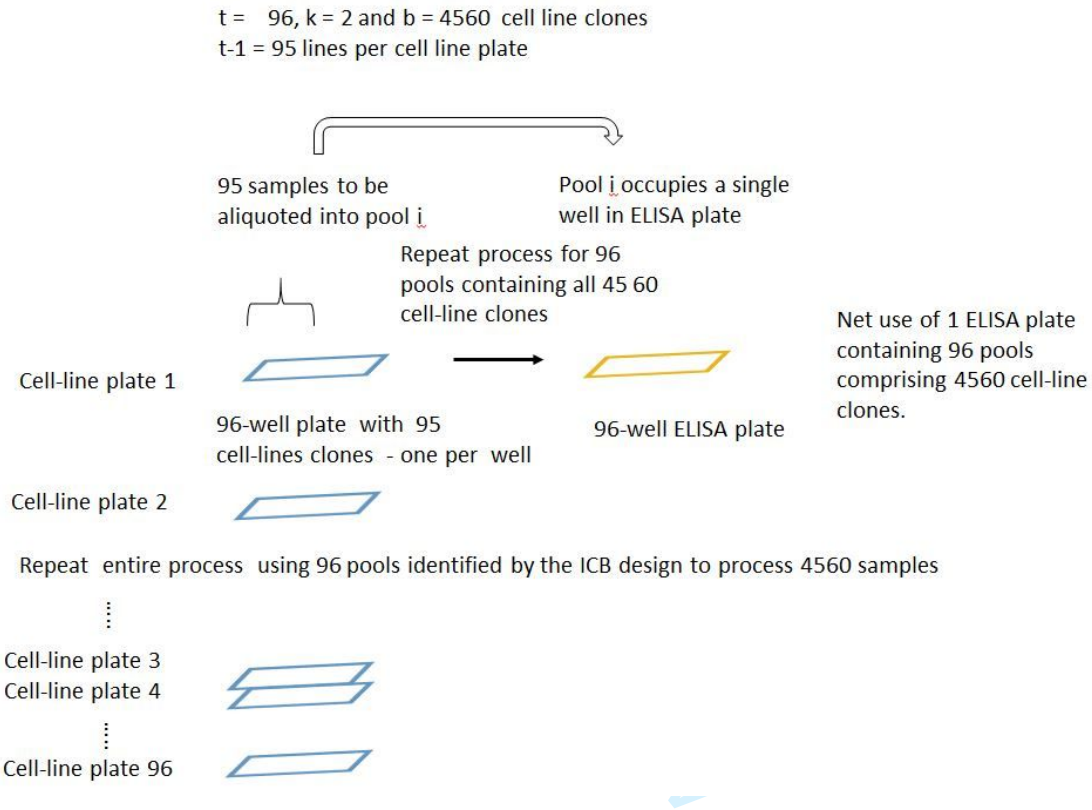

Figure 1. Scenario where dilution effects are negligible.

Set up an ICB-fixed design with t = 96, k = 2 and b = 4560 cell-line clones (specimens). The initial number of pools for the first stage would be 97 (i.e., t + *1*) which implies that first a single pool that contains all clones tests positive. This of course assumes the single pool that contains all clones is actually feasible, but it is not absolutely necessary. The expected number of tests after the first 97 pools is a function of prevalence (*1*-q) of the specific antibody in the population being tested. Positive clones indicated from the first step

could be tested individually, as assumed by ICB-F fixed design or setup in another ICB-S sequential design in order to continue to minimize costs. In this example, the ICB-F design has expected sample size of 99 pools which is considerably smaller than the expected samples sizes of the Dorfman, 2-way and 3-way which are 288, 157 and 145, respectively. The expected savings are approximately 98% of the cost of ELISA screening reagents (e.g., one ELISA plate for screening 96 pools *vs* 48 96 well ELISA plates if the researcher were to screen individual samples). We chose to omit the need for positive and negative controls in the screening plate for simplicity and focused instead on delivering the main point, applicability.

The example described in Figure 1 serves several purposes. First, the scenario is outside of the options listed in Table 2. That is, the use of pooling designs described above is by no means limited to the parameters shown on Table 2. ICB and other designs can be applied to any number of samples (emphasis on the word "any"). Table 2 has the main purpose of comparing design efficiencies under the same parameters and no one design is superior all the time, although, all designs offer different degrees of resource savings compared to the option of individual specimen testing (except for the 1-pool design in some instances). Second, the use of $b = 4560$ clones above serves the purpose of exemplifying a number that fits common sample-handling platforms (i.e., 96-well plates) to ease handling of samples. This approach can be extended to accommodate 384 well-plates, in which case $t = 384$, $k = 2$ and $b = 73{,}536$ results in a scenario where a researcher could screen a total of 73,536 clones using only $t + 1$ (385) pools. Here all pools, other than the first pool that contains all specimens, would contain 383 specimens, and ultimately each pool would be

assigned to a well on a single 384-well ELISA plate (if screening *t* pools only). In other words, these designs are flexible and could easily be scaled up.

Comparison with other designs

The goal of most group screening programs is to identify the positive specimens using a minimum number of pools. In comparing designs, we assume that cost is of primary importance and that the preferred design has the smallest expected sample size (i.e., the smallest expected number of retests + initial number of pools). Two types of ICB group screening designs will be used: (1) fixed designs where retested specimens are tested separately and (2) sequential designs where retested specimens are tested in a single follow-up ICB group screening design. For the sequential designs, the expected sample sizes for the follow-up design will be based on the minimum possible q that could yield a given number of – signs in the response vector of the initial design. For example, for an initial design with $t=11$, $k=5$ where there are $p=2$ – signs in the response vector, the minimum number of positive specimens that could give such a response is 2 (9/5 or $(t-p)/k$ rounded up) giving an estimated $q = 1 - 2/462$ where there are $b = \binom{11}{5} = 462$ specimens to be tested. This q is then used to estimate the expected sample size in the $t=9$, $k=5$ ICB follow-up design. The multi-stage approach with multiple successive follow-up ICB designs, was initially considered, but expected sample sizes differed only slightly from the sequential designs (data not shown) and was not be included.

A number of different group screening schemes have been proposed in the literature where most have been based on either the one-way (*4*) or two-way array designs. (*16*), (*17*), (*5*), (*18*) An extension of the two-way design to a multi-way array (*8*) shows promise and

will be considered. Also, since q is assumed large, a single pool of all the specimens is the simplest possible group screening design, since if the test for this single pool is negative, no further testing is needed. The expected sample sizes will be determined for the two types of ICB designs (fixed: ICB-F, and sequential: ICB-S) the Dorfman approach with optimal pool size as described in Feller, (*19*) (page 240, $k=(1-q)^{-.5}$), the two-way array (*5*), the three-way array (3P method from Berger et al. (*8*)) and the single pool, where all specimens are tested individually if the single, all specimen pool is positive. The pool sizes for the two-way and three-way array designs were approximated as $b^{1/2}$ and $b^{1/3}$ respectively, to enable direct comparison with the ICB designs.

Expected sample sizes for all the methods for a range of specimen sizes with q ranging from 0.95 to 0.999 and an initial number of ICB pools of 8, 12, 16 and 20 (including the single pool) are given in Tables 2. With all methods, the expected number of tests decreased as q increased showing increased efficiency as the attribute became rarer. The single initial pool design nearly always had the largest expected sample size indicating that when a positive event is rare (q _ 0.95), this method can be expected to be the most costly among the methods evaluated. The ICB fixed and sequential ICB designs were nearly identical in most cases meaning for these values of q, the sequential approach had little effect. In a majority of the cases, the Dorfman design had the smallest expected sample size which was based on Feller (*19*) (p240, $2b\sqrt{(1-q)}$). However, for the middle ranges of the number of specimens (b) for a given q (e.g. q=0.95, $55 \le b \le 171$; q=0.99, $105 \le b \le 455$; q=0.999, $330 \le b \le 3876$), the ICB design had the smallest sample size. For the most part, the 2-way and 3-way designs did not perform as well, yet there were a few cases where the 2-way did best and one where the 3-way did best.

The Dorfman, two-way and three-way designs have been compared in previous work but the ICB designs proposed here have not been compared with previous designs. Berger et al. (*8*) evaluated efficiency of two- and three-way arrays and compared efficiencies to upper bounds on multi-stage strategies and showed the three-way array to be more efficient than the two-way for large values of q. Phatarfod and Sudbury (*5*) compared efficiencies of the Dorfman approach, the two-way array and a modified two-way array and concluded that the two-way array was more efficient than the Dorfman approach in all cases considered. Our results were similar to Berger et al. (*8*) for q=0.999 in that the three-way array was more efficient than the two-way in the majority of cases, but we found the 2-way method more efficient than the 3-way in the majority of the remaining cases. However, contrary to Phatafod and Sudbury, (*5*) we found the Dorfman design to be more efficient than the two-array. The reason our results differed was likely because we did not 'optimize' pool size for the two- and three-way arrays. Optimizing pool size for two- and three-way array designs is often not appropriate since in most screening programs, the number of specimens to be screened is known before work begins as opposed to identifying the number of specimens to be screened as a function of q. In addition, assuming b and q are given allowed direct comparison with the ICB designs and the other methods.

In most applications, the number of specimens (b) will be known and the choice of the number of pools (t) (excluding the all specimen pool) and the number of specimens per pool $\binom{t-1}{k-1}$ will be of interest. When b is not in Table 2, a reasonable design can be found. As a specific example, assume there are 80 specimens to be evaluated for a prevalence of 0.05 (*1-q*) and the intent is to determine approximate values of the number of pools and the number of specimens per pool for the initial design. From Table 2, with q=0.95, there is no

b=80 ICB design, but there are ICB-F designs for $b=55$ and $b=105$ with expected sample sizes of 15 and 20 with 10 and 14 specimens per pool. A reasonable approximation of the number of pools (t) is 13 which may be found by finding t such that $80 \cong \binom{t}{2} = 78$. The precise initial design would have 78 specimens in an ICB design with 13 pools (t=13), 12 specimens per pool ($\binom{13-1}{2-1}$=12) and an additional 2 specimen pool of the 79th and 80th specimens. This design will have an approximate expected number of pools to be 17.5 (1/2 way between 15 and 20). As a comparison with the other designs: (1) the optimal Dorfman design will have $k \cong 5$ with $t=16$ and with an overall expected sample size of 36 (specifically, from Feller (1968), $k \cong 1/\sqrt{(1-q)} = 1/.2235 \cong 5$; t=80/5=16 and expected sample size of $2b*.2235=160*.2235=36$); (2) the approximately optimal 2-way array design will have 9 pools with 9 specimens each, ie $9 \cong b^{1/2}$ where b=80 with an interpolated expected number of samples of 29; (3) the approximately optimal 3-way array will have $k=80^{1/3} \cong 4$, (ie 4-4x4 arrays) with a reasonable initial design of a 5 4x4 3-way array with an interpolated expected number of samples of 38. In the three designs being compared to the ICB-F design, each has a larger expected sample size, some by a considerable margin.

The same type of approach can be used to obtain reasonable ICB designs for all values of b that fall within the range of Table 2. For a b value that is between two b values from Table 2 that have the ICB-F designs with the smallest expected sample sizes, the approximate ICB-F design as described above should be more efficient than any of the other designs. For example, when q=0.999 and $330 \le b \le 3876$, the approximate ICB-F design would appear to be preferred since the expected sample size at b=330 is 12 and at b=3876, the expected sample size is 56, both of which are smaller than all other comparable designs.

But it must be remembered that all of the above comparisons are only approximations and further research into the properties of these designs beyond those given in Table 2 is needed.

A major assumption is that there are no testing errors. ICB group screening designs may be still useful in the presence of one false negative or one false positive pool in the initial experiment, although more testing likely will be necessary. If it is unknown that an *r* specimen pool is a false negative and $k+j'$ ($j'=1,\ldots,t-k+1$) positive pools are observed, then the response vector will make it appear that $\binom{t-p}{k}=\binom{k+j'-1}{k}$ specific specimens should be retested. However, upon retesting these specific specimens, none will be positive, and it can be shown that another set of $\binom{k+j'-1}{k-1}(t-k-(j'-1))$ specimens should be retested thus increasing the expected sample size. Similarly, if it is unknown that an *r* specimen pool is a false positive and $k+j'$ ($j'=0,\ldots,t-k+1$) positive pools are observed, it will appear that a set of $\binom{t-p}{k}=\binom{k+j'-1}{k}$ specific specimens should be retested, however only specimens from a subset of $\binom{k+j'-2}{k}$ specimens can truly be contributing to the response vector. Thus, a single false positive in the initial experiment will increase retesting and increase the expected sample size. Similar results can be established for more than one false positive or negative, but the important issue is that the cost savings ability of the ICB group screening designs decreases as testing error increases. Du and Hwang (*15*) (Theorem 3.1.2) also demonstrate that when there is at most one positive specimen, an incomplete block group screening design can tolerate a single false positive or false negative pool.

When it is important to screen a large number of specimens and the chances are small (<0.001) that any one specimen is positive for the attribute of interest, the three-way array and the ICB sequential designs show a considerable cost savings over the two most commonly used group screening designs: the Dorfman approach and the two-way array. The ICB sequential design can give nearly a 90% and 80% savings over the Dorfman and the three-way array approaches respectively and up to a 70% savings over the three-way array approach in some cases. In other cases, the three-way design can show large savings over the Dorfman, two-way and ICB designs. Du and Hwang (*15*) give extensive description of the properties of two-way and three-way designs.

Discussion

We chose to highlight a relatively simple scenario with the example of Figure 1 to hopefully deliver a clear picture of the potential efficiencies and cost savings of these designs. However, we are well aware of the several (sometimes unrealistic) assumptions used. Perhaps, the most critical assumption is the absence of dilution effects when using pools of hundreds of samples. The power of these designs will depend heavily on the robustness of technology available for screening and there will be some instances where threshold parameters will have to be experimentally defined along with variance quantification. It is likely that before adopting an efficient pool screening design methodology, a lab may need to invest resources to identify technology limitations and quantify technical variation. However, such investments can have large payoffs since these designs can yield great efficiencies and cost savings and more importantly are backed up by a sound statistical framework. When using ICB designs, the randomization of samples as

well as the actual pool design to be used by the bench scientist can be easily obtained with an R program (see appendix).

The use of pooling designs can extend to other areas of research. Discovery of rare functional variants: Group screening designs can be used to decrease the cost of sequencing runs in experiments estimating the impact of minor allele frequency (MAF) on detection capability of functional variants when using pools of individuals. Jakaitiene et al. (*20*) showed a novel methodology using pooled patient samples that allowed an efficient increase in sample size by sequencing pools of individuals. This, tied together with their clever multi-reference (multi-pool as reference) framework and beta-binomial models allowed them to improve the accuracy of identification of neuromuscular disorder mutations, and it also allowed for the discovery of novel rare variants (Jakaitiene et al. (*20*)). Their multi-reference framework can be further extended to make use of ICB group screening designs to increase the power of the reference pools by determining the optimal number of pools to be used in an experiment, so as to maximize efficiency and detection capabilities. Groups of pools using ICB group screening designs can also be partitioned across multiple sequencing runs and the independent runs can be analyzed together as a meta-analysis (Wang et al. (*21*)). The potential for ICB group screening designs in the detection of rare causative disease alleles can have a huge impact in the field of epidemiology where the screening of vast number of samples is the nature of this field (Wang et al. (*21*)).

Recent discoveries in gene editing (CRISPR-Cas technology) are potential tools adapted to on-field deployable diagnostic tests. The CRISPR-Cas gene editing system has been further modified to allow for the rapid identification of very specific viral infections in samples that resemble real life human samples (Chen et al. (*22*); Chertow (*23*); Gootenberg

et al. (*24*); Myhrvold et al. (*25*)). The Cas13 system adapted into the SHERLOCK (specific high-sensitivity enzymatic reporter unlocking) platform, together with HUDSON (heating unextracted diagnostic samples to obliterate nucleases, an approach to make detection of nucleic acids rapid and direct) have been shown to hold great promise for the rapid, precise and effective identification of infected patients with single or multiple virus biotypes (Myhrvold et al. (*25*)). The key characteristics of these optimized systems are the instrument-free and fast-paced methods possible. The authors engage in a discussion that highlights the use of this technology in parts of the world where diagnostic testing is often cost-prohibitive and not used (Chen et al. (*22*); Chertow, (*23*); Gootenberg et al. (*24*); Myhrvold et al. (*25*)). If such a postulate is true and if we are to strive to make new technologies available to underdeveloped regions in the world, then we must seriously explore the possibilities of incorporating efficient group screening designs, based on sound statistical foundations, in order to substantially bring down the per-sample cost to truly realize this technology's promise.

Conclusion

Rapid and multiplexed diagnosis and screening tests can be productively applied in a multitude of areas such as drug discovery, precision oncology, cancer immunotherapy, gene editing, high-throughput sequencing, government funded test screening and large scale genetic interaction studies to name a few. The search for the "needle in the haystack" is the rule and not the exception in many scientific disciplines. The hunt of rare events has prompted a special focus to the dissemination of knowledge in drug development and related areas and we hope this work contributes to discussion (Domach (*26*)). It is clear that there are many technical challenges of current screening platforms and consequently, sound

statistical experimental design becomes imperative and not a mere formal nuance. Group screening strategies to target a low frequency (rare) event, can be productively applied to many of these areas with important impacts on areas ranging from agriculture to biosecurity and human health. The different group screening designs proposed above assume that variation in the specific technique can be controlled to a reasonable level resulting in adequate sensitivity and specificity. It is important as we move forward with technology dictating the research that we conduct, to intelligently use efficient and precise statistical tools such as group screening designs.

Acknowledgement

The authors would like to acknowledge some early discussions on this work with the late Neil Butler, who passed away on 21 June 2009. Neil was an exceptionally imaginative researcher in the design of experiments and had published extensively on the subject. He is greatly missed by his friends and colleagues.

Literature Cited

1. Meyerson, L. A.; Reaser, J. K. Biosecurity: Moving toward a Comprehensive Approach: A comprehensive approach to biosecurity is necessary to minimize the risk of harm caused by non-native organisms to agriculture, the economy, the environment, and human health. *AIBS Bulletin* **2002,** *52,* 593-600.
2. Zhang, M.; Chen, W.; Chen, X.; Zhang, Y.; Lin, X.; Wu, Z.; Li, M. Multiplex immunoassays of plant viruses based on functionalized upconversion nanoparticles coupled with immunomagnetic separation. *Journal of Nanomaterials* **2013,** *2013,* 122.
3. Johnson, M. Screening Designs. *Encyclopedia of Statistical Sciences* **2004,** *9*.
4. Dorfman, R. The detection of defective members of large populations. *The Annals of Mathematical Statistics* **1943,** *14,* 436-440.
5. Phartarfod, R. M.; Sudbury, A. The use of a square array scheme in blood testing. *Statistics in Medicine* **1994,** *13* (22), 2337-2343.
6. Hudgens, M. G.; Kim, H.-Y. Optimal configuration of a square array group testing algorithm. *Communications in Statistics—Theory and Methods* **2011,** *40,* 436-448.
7. Sudbury, A. Two-stage testing using selection schemes. *Statistics in medicine* **2010,** *29,* 2194-2199.
8. Berger, T.; Mandell, J. W.; Subrahmanya, P. Maximally efficient two-stage screening. *Biometrics* **2000,** *56,* 833-840.
9. Kim, H.; Hudgens, M.. Three-dimensional array-based group testing algorithms. *Three-dimensional array-based group testing algorithms* **2009,** *65* (3), 903-910.
10. Sterrett, A. On the detection of defective members of large populations. *The Annals of Mathematical Statistics* **1957,** *28,* 1033-1036.
11. McMahan, C. S.; Tebbs, J. M.; Bilder, C. R. Informative Dorfman screening. *Biometrics* **2012,** *68,* 287-296.
12. Redman, C. E.; King, E. P. Group screening utilizing balanced and partially balanced incomplete block designs. *Biometrics* **1965,** 865-874.
13. Bush, K. A.; Federer, W. T.; Pesotan, H.; Raghavarao, D.; others. New combinatorial designs and their applications to group testing. *Cornell Biometrics Unit Technical Reports* **1980,** 1-15.
14. Balding, D. J.; Torney, D. C. The design of pooling experiments for screening a clone map. *Fungal Genetics and Biology* **1997,** *21,* 302-307.
15. Du D, H. F. K. Pooling designs and nonadaptive group testing. *Important tools for DNA*

sequencing. Series on applied mathematics **2006,** *18*.

16. Evans, G. A.; Lewis, K. A. Physical mapping of complex genomes by cosmid multiplex analysis. *Proceedings of the National Academy of Sciences* **1989,** *86,* 5030-5034.
17. Zwaal, R. R.; Broeks, A.; Meurs, J.; Groenen, J. T.; Plasterk, R. H. Target-selected gene inactivation in Caenorhabditis elegans by using a frozen transposon insertion mutant bank. *Proceedings of the National Academy of Sciences* **1993,** *90,* 7431-7435.
18. Cai, W.-W.; Chen, R.; Gibbs, R. A.; Bradley, A. A clone-array pooled shotgun strategy for sequencing large genomes. *Genome Research* **2001,** *11,* 1619-1623.
19. Feller, W. *An introduction to probability theory and its applications;* Wiley, New York, 1968; Vol. 1.
20. Jakaitiene, A.; Avino, M.; Guarracino, M. R. Beta-Binomial Model for the Detection of Rare Mutations in Pooled Next-Generation Sequencing Experiments. *Journal of Computational Biology* **2017,** *24,* 357-367.
21. Wang, Q.; Lu, Q.; Zhao, H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in genetics* **2015,** *6,* 149.
22. Chen, J. S.; Ma, E.; Harrington, L. B.; Da Costa, M.; Tian, X.; Palefsky, J. M.; Doudna, J. A. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **2018,** *360,* 436-439.
23. Chertow, D. S. Next-generation diagnostics with CRISPR. *science* **2018,** *360,* 381-382.
24. Gootenberg, J. S.; Abudayyeh, O. O.; Kellner, M. J.; Joung, J.; Collins, J. J.; Zhang, F. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **2018,** *360,* 439-444.
25. Myhrvold, C.; Freije, C. A.; Gootenberg, J. S.; Abudayyeh, O. O.; Metsky, H. C.; Durbin, A. F.; Kellner, M. J.; Tan, A. L.; Paul, L. M.; Parham, L. A.; others. Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **2018,** *360,* 444-448.
26. Domach, M. Finding the proverbial needle in the haystack… & then what? Clonally-Derived cell lines. *Biotechnology progress* **2018,** *34,* 557-557.
27. Koepf, W. *Hypergeometric summation. An algorithmic approach to summation and special function identities, Universitext;* Springer, London, 2014.

Table 1. Small ICB group screening design with 10 specimens and 6 pools

Pool	Specimen									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	+	+	+	+						
2	+				+	+	+			
3		+			+			+	+	
4			+			+		+		+
5				+			+		+	+
6	+	+	+	+	+	+	+	+	+	+

Table 2. Expected sample sizes for six group screening designs with the probability that a specimen fails to contain the attribute of interest (*q*) of 0.95, 0.99, 0.999, the number of specimens tested (b), an initial number of pools in the ICB design (i_pools) of *8, 12, 16, 20* and specimens per pool (spp).

q	b	i_pools	spp	t	k	ICB-S	ICB-F	Dorf	2-way	3-way	1-pool
0.95	21	8	6	7	2	8	9	9	9	10	14
0.95	21	8	15	7	5	12	12	9	9	10	14
0.95	35	8	15	7	3	11	13	16	12	15	29
0.95	35	8	20	7	4	15	16	16	12	15	29
0.95	55	12	10	11	2	13	15	25	20	23	52
0.95	105	16	14	15	2	18	22	47	38	53	105
0.95	165	12	45	11	3	79	92	74	57	102	165
0.95	171	20	18	19	2	22	33	76	68	108	171
0.95	330	12	120	11	4	335	336	148	156	275	330
0.95	455	16	91	15	3	414	424	203	236	417	455
0.95	462	12	210	11	5	474	474	207	236	425	462
0.95	969	20	153	19	3	928	984	433	673	980	969
0.95	1365	16	364	15	4	1365	1381	610	993	1391	1365
0.95	3003	16	1001	15	5	3003	2999	1343	2671	3046	3003
0.95	3876	20	816	19	4	3876	3876	1733	3655	3923	3876
0.95	11628	20	3060	19	5	11628	11628	5200	11569	11696	11628
0.99	21	8	6	7	2	8	8	4	8	9	5
0.99	21	8	15	7	5	8	8	4	8	9	5
0.99	35	8	15	7	3	8	8	7	10	10	11
0.99	35	8	20	7	4	9	9	7	10	10	11
0.99	55	12	10	11	2	12	12	11	15	12	24
0.99	105	16	14	15	2	16	17	21	22	17	68
0.99	165	12	45	11	3	13	14	33	27	24	134
0.99	171	20	18	19	2	21	21	34	30	24	140
0.99	330	12	120	11	4	48	56	66	47	49	318
0.99	455	16	91	15	3	22	32	91	61	78	450
0.99	462	12	210	11	5	210	222	92	61	80	458
0.99	969	20	153	19	3	61	124	194	136	293	969
0.99	1365	16	364	15	4	998	1038	273	198	549	1365
0.99	3003	16	1001	15	5	2973	3017	601	629	2084	3003
0.99	3876	20	816	19	4	2403	3879	775	962	3041	3876
0.99	11628	20	3060	19	5	11628	8766	2326	5197	11497	11628
0.999	21	8	6	7	2	8	8	1	8	8	1
0.999	21	8	15	7	5	8	8	1	8	8	1
0.999	35	8	15	7	3	8	8	2	10	10	2
0.999	35	8	20	7	4	8	8	2	10	10	2
0.999	55	12	10	11	2	12	12	3	14	11	4

0.999	105	16	14	15	2	16	16	7	20	14	11
0.999	165	12	45	11	3	12	12	10	24	17	26
0.999	171	20	18	19	2	20	20	11	26	17	28
0.999	330	12	120	11	4	12	12	21	36	21	94
0.999	455	16	91	15	3	16	16	29	43	24	167
0.999	462	12	210	11	5	13	13	29	43	24	172
0.999	969	20	153	19	3	21	21	61	64	33	602
0.999	1365	16	364	15	4	18	18	86	75	40	1017
0.999	3003	16	1001	15	5	118	144	190	119	80	2854
0.999	3876	20	816	19	4	32	56	245	141	113	3796
0.999	11628	20	3060	19	5	5913	6257	735	341	935	11628

t=no treatments in incomplete block; k=block size; ICB-S =expected sample size using unreduced ICB with sequential first follow-up using unreduced incomplete block (ICB) design; ICB-F =expected sample size using fixed unreduced incomplete block (ICB) design; Dorf. = Dorfman expected sample size using Dorfman one-way design using optimal pool size from Feller (1968); 2-way = exp. sample size using two-way square array where a side =square root (b); 3-way = exp. sample size using Berger et al. 3-way array design with pool size = b** (1/3); 1-Pool = exp. sample size using a single pool and testing each specimen if pool is positive.

Appendix

Derivation of C_{u1} and/or C_{uv} coefficients

Proof: The C_{u1} coefficients of the $q^{b-2}(1-q)^2$ terms are obtained by considering pairs of blocks and determining how many pairs are possible for a given number of – signs in the response vector. For any particular block that is to be paired with another block, there are $\binom{k}{j'}\binom{t-k}{k-j'}$ different possible pairs of blocks with $j'(j'=0,\ldots,k-1)$ treatments in common in both blocks meaning there will be $t+j'-2k$ – signs in the response vector. The total number of different possible pairs for a given number of common treatments is obtained by multiplying $\binom{k}{j'}\binom{t-k}{k-j'}$ by $\binom{t}{k}\div 2$ since there are $\binom{t}{k}$ blocks where each pair would be counted twice without division by 2.

The C_{uv} $(v\ 1)$ coefficients are based on an extension of the concept from the C_{u1} coefficients and are developed as recursive functions of $C_{u,v-1}$. For any given set of $v-1$ blocks with k' treatments ($k'\le k$) present in the set, there will be $\binom{k'}{j'}\binom{t-k'}{k-j'}$ different possible sets of v blocks with j' treatments in common between the added block and the given set of $v-1$ blocks, except when all j' treatments are from the k' treatments (i.e. $j'=k$), in which case, the number of different possible sets will be reduced by $v-1$ since $v-1$ blocks are given. Under these conditions, there will be $t+j'-k'-k$ – signs in the response vector. By relabeling indices and summing, the Vandermonde convolution (*27*) gives $\sum_{u=j}^{k+j}\binom{k+j}{k+j-u}\binom{t-(k+j)}{k-(k+j-u)}=\binom{t}{k}$ for $k=1,\ldots,t-1$ and $j=1,\ldots,t-k$. Thus, summing u from j

to $k+j$ over the different possible sets of v blocks is $\binom{t}{k}$ - *(v-1)*. A recursive relationship between combinations is $\binom{b}{v}=\binom{b}{v-1}\left(\frac{b-(v-1)}{v}\right)$ where $b=\binom{t}{k}$ but $\sum_{j=1}^{t-k} C_{j,v-2}=\binom{b}{v-1}$ since all coefficients must sum to the appropriate binomial coefficients and so $\binom{b}{v}=\left(\sum_{j=1}^{t-k} C_{j,v-2}\right)\left(\sum_{u=j}^{k+j}\binom{k+j}{k+j-u}\binom{t-(k+j)}{k-(k+j-u)}-(v-1)I(j=u)\right)\div v$. Finally, since

$$\binom{b}{v+1}=\sum_{j=1}^{t-k} C_{j,v}\ ,\ C_{uv}=\sum_{j=1}^{k+1} C_{j,v-1}\left[\binom{k+j}{j+k-u}\binom{t-(k+j)}{k-(j+k-u)}-vI(j=u)\right]\div(v+1).$$

R Program to group specimens to pools

```
## install.packages("crossdes")
## install.packages("ibd")
## run under R version 3.4.4
library(crossdes)
library(ibd)

## Enter number of specimens below (use real integers)
NumberOfSpecimens <- 10

## randomize specimens. We assume your specimens are labeled with letters
randomSpecimens <- rbind ( sample ( seq ( 1 : NumberOfSpecimens ) , size =
NumberOfSpecimens , replace = F , prob = NULL ) , seq ( 1 : NumberOfSpecimens ) )
rownames ( randomSpecimens ) <- c ( "your_specimen_ID" ,
"number_ID_assigned_to_Pool_Matrix" )
randomSpecimens
## example: your specimen (1) in row 1 will be assigned to corresponding number in row 2
given by (randomSpecimens)... and so on
## row2 numbers correspond to numbers in (Pool_Matrix) below

## the example below refers to: 5 choose 2 = NumberOfSpecimens
A = find.BIB ( 5 , NumberOfSpecimens , 2 )
N = design_to_N ( A )
numRows <- dim ( N ) [ 1 ]
numCols <- dim ( N ) [ 2 ]
colnames ( N ) <- seq ( 1 , numCols )
rownames ( N ) <- seq ( 1 , numRows )
specimensAll <- NULL
```

```
for ( i in 1 : dim ( N ) [ 1 ] ) {
  temp1 <- N [ i , N [ i , ] %in% c ( 1 ) ]
  specimensAll <- rbind ( specimensAll , names ( temp1 ) )
}
Pool_Matrix <- t ( specimensAll )
colnames ( Pool_Matrix ) <- paste ( "Pool" , seq ( 1 , dim ( Pool_Matrix ) [ 2 ] ) , sep = "_" )
Pool_Matrix
```